Parametric encoder and method for encoding an audio or speech signal

The invention relates to a parametric encoder and method for encoding an audio or speech signal into sinusoidal code data.

Such encoders and methods are generally known in the art and are for example disclosed in B. Edler, H. Purnhagen, and C. Ferekidis "ASAC - Analysis/synthesis codec for very low bit rates", Preprint 4179 (F-6) 100th AES Convention, Copenhagen, 11-14 May 1996. Such a known parametric encoder is illustrated in Figs. 4 and 5.

According to Fig. 5 the encoder comprises a segmentation unit 120' for segmenting a received audio or speech signal into at least one single scale segment $x_m(l)$ having the samples $x_m(0)$, ..., $x_m(L-1)$. These samples are received by a sinusoidal estimation unit 140', for estimating sinusoidal code data representing said segment $x_m(n)$. These sinusoidal code data are typically merged into a data stream before been transmitted via a channel or stored on a recording medium.

Fig. 4 provides an - also known - more detailed illustration of the segmentation unit 120'. As can be seen there, the audio or speech signal s(n) is input into a tapped delay line comprising consecutive filters 122_1', 122_2', ..., 122_L-1'. The original audio or speech signal $s(n) = y_0(nD)$ as well as the output signals $y'_1(nD)$ ..., $y_{L-1}(nD)$ of said L-1 filters 122_1', ... 122_L-1' are input into a sampling unit 124', preferably embodied as down sampling unit, in order to generate L samples $x_m(0)$, ..., $x_m(L-1)$ of the segment $x_m(l)$.

The single scale segments as generated by the known parametric encoder according to Figs. 4 and 5 are characterised in that their segment length and consequently also their frequency resolution is constant independent of the actual frequency range of the segmented audio or speech signal. Expressed in other words, the single scale sinusoidal estimation mechanism as provide in the common encoders gives problems with the required time-frequency resolution trade-off. In particular for low frequency ranges of the signal s for

high-quality audio coding high frequency resolution is required, whereas for other frequency ranges a lower frequency resolution, i.e. a lower segment length L would be sufficient.

In order to overcome these problems, multi-scale models have been proposed, for example by T.S. Verma S.N. Levine and J.O. Smith III "Multiresolution sinusoidal modeling for wideband audio with modifications", in Proc. ICASSP-98, Seattle, 1998. These multi-scale models provide different segment length L for different frequency ranges of the signal s. However, these multi-scale models bring about problems of scattering of components over scales and/or of merging the data retrieved at different scales. More specifically, a problem of scattering addresses the problem that the generated segments usually overlap and thus, samples of said segments might be processed twice because there is no clear separation possible - except of applying high effort - between the samples of two generated segments.

Starting from that prior art it is an object of the invention to improve a known parametric encoder and method for encoding an audio or speech signal such that a required time-frequency resolution trade-off can be established without having the above mentioned problems of the multi-scale models, namely the problem of scattering of components over scales and/or of merging the data retrieved at different scales.

This object is solved by the subject matter of claim 1. More specifically, for the known parametric encoder it is suggested according to claim 1, that the segmentation unit is further embodied for carrying out a frequency-warping operation in order to transform the output samples onto a frequency-warped domain and to provide a post-processing filter for re-mapping said sinusoidal code data output from the sinusoidal estimation unit to the original frequency domain of the signal s.

The segmentation unit of the claimed parametric encoder segments the signal s into at least one single scale segment $x_m(l)$. Because said unit only generates single scale segments the problems of the multi-scale models known in the art do not occur here. Instead, by applying the frequency-warping operation the required time-frequency resolution trade-off, i.e. providing different frequency resolutions for different frequency ranges of the signal s, can advantageously be established for single scale segments without any problems.

It shall be noted here that unilateral frequency-warping is generally known in the art, e.g. for linear predictive coding of audio, audio equalisation and by normal filter design, but not for sinusoidal coding as suggested in that application. Bilateral frequency warping has not been applied in audio processing.

Advantageous embodiments of that parametric encoder are mentioned in the dependent claims.

The object is further solved by a method for encoding an audio or speech signal according to claim 9. The advantages of said method correspond to the advantages mentioned above for the parametric encoder.

Five figures are accompanying the description, wherein

Fig. 1 shows a first preferred embodiment of the parametric encoder according to the invention;

Fig. 2 shows a second preferred embodiment of the parametric encoder according to the invention;

Fig. 3 shows a third preferred embodiment of the parametric encoder according to the invention;

Fig. 4 shows a detailed illustration of a parametric encoder known in the art; and

Fig. 5 shows a general block diagram of the parametric encoder known in the art.

In the following the preferred embodiments of the parametric encoder according to the invention are described by referring to Figs. 1 to 3.

Fig. 1 shows a first preferred embodiment of the parametric encoder according to the invention for encoding an audio or speech signal s(n) into sinusoidal code data scd. It comprises a segmentation unit 120 for segmenting said signal s into at least one single scale segment $x_m(n)$ with m = 1 ... M, where m denotes a current downsampling step. More specifically, said segmentation unit 120 comprises a plurality of L-1 filters 122_1, ..., 122_L-1 being connected in series for receiving the signal s(n) at the input of the first of said filters

122_1. Said segmentation unit 120 further comprises a sampling unit 124 for receiving and preferably down sampling said signal $s(n) = y_0(n)$ as well as the output signals $y_1(n)$ ..., $y_{L-1}(n)$ of said L-1 filters 122_1, ..., 122_L-1 in order to generate L samples $x_m(0)$, ..., $x_m(L-1)$ of the single scale segment $x_m(l)$ with $l = 0$ ...(L-1). In said first embodiment all of the L-1 filters 122_1, ..., 122_L-1 are embodied as all-pass filters having a transfer function A(z) defined as:

$$A(z) = \frac{-\lambda^* + z^{-1}}{1 - \lambda z^{-1}}, \tag{1}$$

where * denotes a complex-conjugation and $|\lambda| < 1$. Typically, λ is real-valued and $\lambda \neq 0$.

In that first embodiment the processing is the following:

The audio signal s is input to a tapped all-pass line having outputs $y_l(n)$ ($l = 0,1, ..., L-1$) with

$$y_0(n) = s(n), \text{ and} \tag{2}$$

$$y_l = y_{l-1} * \alpha \text{ for } l = 1,2, ..., L-1 \tag{3}$$

with * denoting convolution and α the impulse response associated with the transfer function A(z). The outputs y_l are downsampled (read-out every D time instances) and defined as a segment x_m:

$$x_m(l) = y_l(mD) \tag{4}$$

where D is the downsampling factor of the sampling unit 140. The signal output by said sampling unit 124 is considered to represent the samples $x_m(l)$ with $l = 0, 1, ..., L-1$ of a segment x_m.

It is important to note that because the filters 122_1, ..., 122_L-1 are - according to the first embodiment -embodied as all-pass filters the samples output by the sampling unit 124 are on a frequency-warped domain.

Said samples $x_m(l)$ with $l = 0, ..., L-1$ are input into a sinusoidal estimation unit 140 for estimating the sinusoidal code data representing the segment x_m. The estimation may

be done by carrying out a Fourier transformation on said frequency-warped samples and subsequent, for instance, peak picking.

It is further important to note that the sinusoidal code data as output by said sinusoidal estimation 140 is on a frequency-warped domain. Consequently, said sinusoidal code data has to be re-mapped, i.e. to be de-warped, to the original frequency domain of the audio or speech signal s. This is done by a post-processing filter 160 following said sinusoidal estimation unit 140. The output of said post-processing filter 160 corresponds to the re-mapped sinusoidal code data associated with the original signal segment x_m.

After sinusoidal extraction, as finished by said post-processing filter 160, the subsequent processing step is residual modelling. The cheapest way of residual modelling is using a parametric model for the power spectral density functions. Such an approach allows the integration of sinusoidal and noise estimation since, for noise modelling frequency-warping can be used.

In the first embodiment the frequency warped samples warped by said sampling unit 120 belong to a single scale segment x_m with the result that the problems of multi-scale models known in the art do not occur here. Due to the embodiment of the filters as all-pass filters a frequency-warping operation is carried out resulting in the frequency-warped samples at the output of the sampling unit 124. Due to the frequency warping operation the required time-frequency resolution trade-off is achieved for the signal s. However, disadvantageously, the power spectral density function of the original audio or speech signal is slightly amended.

Fig. 2 shows a second embodiment of the parametric encoder which substantially corresponds to the first embodiment. In particular, the sampling unit 124, the sinusoidal estimation unit 140 and the post-processing filter 160 in the second embodiment are identical to the corresponding units in the first embodiment. Moreover, the filters 122_3, ..., 122_L-1 correspond to the respective filters in the first embodiment because they are also embodied as first-order all-pass filters having a transfer function A(z) according to equation (1).

However, the second embodiment differs from the first embodiment in that the first filter 122_1 in the series connection of filters in the segmentation unit 120 has a transfer function $A_0(z)$ according to:

$$A_0(z) = \frac{1}{1-\lambda z^{-1}}, \tag{5}$$

Moreover, the second filter 122_2 is also not embodied as all-pass filter but has instead a transfer function $A_1(z)$ according to

$$A_1(z) = \sqrt{1-|\lambda|^2}\,\frac{z^{-1}}{1-\lambda z^{-1}}, \tag{6}$$

wherein in equations 5 and 6 λ is typically real-valued.

For $\lambda > 0$ the transfer functions $A_0(z)$ and $A_1(z)$ both represent a low-pass filter, whereas for $\lambda < 0$ both transfer functions represent a high-pass filter.

The advantages of the second embodiment correspond to the first embodiment. Moreover, the shape of the power spectral density function of the original audio or speech signal s is better maintained.

A problem the first and second embodiment is that the introduced frequency warping operation acts as a unilateral device. The past is warped and, as a consequence of the fact that effectively the time-scale for each frequency is different, the estimated frequencies are good estimates for the instantaneous frequencies some n samples ago, where n, representing delays of the instantaneous frequencies, is dependent on the instantaneous frequencies themselves. Expressed in other words, the presence of the delay as such is accepted, but its frequency dependency should be avoided because this frequency dependency is disadvantageous for encoding purposes; for encoding purposes an estimate of the instantaneous frequencies at a well-defined moment in time is desired.

To achieve this, it is proposed to extend the frequency-warping procedure to a bi-lateral operation, warping both, the past and the future. The latter is not possible with the

mechanisms considered in embodiments 1 and 2 since these are based on infinite-impulse response IIR-filters.

However, considering the frequency-warping of a finite segment and observing a finite part of the ideally infinitely-long warped signal then the processing using IIR-filters reduces to a matrix-vector multiplication. In that case the parametric encoder can be embodied according to a third embodiment of the invention as shown in Fig. 3. According to that embodiment the received audio or speech signal is input into a tapped delay line and subsequently said audio or speech signal s as well as the output signals $y_1(n)$ ..., $y_{L-1}(n)$ of the L-1 filters 122_1, ..., 122_L-1 of the tapped delay line are input into a sampling 124 unit for generating a segment x_m having a number of $N_1 + 1 + N_2$ samples being indexed $-N_1, -N_1 + 1$, ..., 0, ..., $N_2 - 1, N_2$ with $N_1, N_2 > 0$. It is important to note that the sampling operation carried out so far in the third embodiment corresponds to the sampling operation known in the art as described by referring to Fig. 4 and that the samples resulting from that common sampling operation at the output of the sampling unit $x_m^0(-N_1), ..., x_m^0(0), ..., x_m^0(N_2)$ are not yet on a frequency-warped domain.

In order to transform the samples onto the frequency-warped domain a bi-lateral warping operation is carried out by an additionally provided bi-lateral warping unit 126, preferably also provided within said sampling unit 120. Said unit carries out the matrix-vector multiplication mentioned in the previous paragraph, written in matrix notation:

$$x_m = Bx_m^0 \tag{7}$$

The transformation matrix B can be calculated for different frequency-warping operations, in particular it can be calculated such that the frequency-warping operations according to embodiment 1 or 2 of the invention are simulated or realised by the third embodiment. The samples output by said bi-lateral warping unit 126 are - in contrast to the input samples - on the desired frequency-warped domain like the samples output by the sampling unit 120 according to embodiments 1 or 2. As can be seen from Fig. 3 the transformed samples are output to the sinusoidal estimation unit 140 in which the desired sinusoidal code data are estimated and finally the sinusoidal code data on the frequency-warped domain is output by said estimation unit 140 and input into the post-processing filter

160 for being re-mapped to the original frequency domain of the signal s. Subsequently, an example for calculating the transformation matrix B is given such that embodiment 2 is simulated by embodiment 3.

In order to achieve this simulation, frequency-warping of a segment $x^0(n)$ having a finite support is considered. More specifically, the samples of said segment are indexed to $-N_1, -N_1+1, \ldots 0, \ldots, N_2$ with $N_1, N_2 > 0$. The associated warped signal is denoted by $\tilde{x}(n)$ and has, in principle, an infinite support.

The Fourier transforms of the sample x(n) and of the associated warped signal are given as

$$S(e^{j\theta}) = \sum_n x(n)e^{-j\theta n}$$

$$\tilde{S}(e^{j\phi}) = \sum_n \tilde{x}(n)e^{-j\phi n}$$

with $j = \sqrt{-1}$. For frequency-warping according to the phase characteristic of an all-pass section the following relation between these frequency variables are given:

$$\phi = \theta + 2 \arctan\left\{\frac{\lambda \sin\theta}{1-\lambda\cos\theta}\right\}, \tag{8}$$

or

$$e^{j\theta} = \frac{e^{j\phi}+\lambda}{1+\lambda e^{j\phi}}. \tag{9}$$

From this it follows that

$$\tilde{x}(n) = \frac{1}{2\pi}\int_{<2\pi>} \tilde{S}(e^{j\phi})e^{j\phi n}d\phi$$

$$= \frac{1}{2\pi} \int_{<2\pi>} S\left(\frac{e^{j\phi} + \lambda}{1 + e^{j\phi}\lambda}\right) e^{j\phi n} d\phi$$

$$= \frac{1}{2\pi} \int_{<2\pi>} \sum_{k=\infty}^{\infty} s(k) \left(\frac{e^{j\phi} + \lambda}{1 + e^{j\phi}\lambda}\right)^{-k} e^{j\phi n} d\phi$$

$$= \sum_{k=-\infty}^{\infty} x(k) \frac{1}{2\pi} \int_{<2\pi>} \left(\frac{e^{j\phi} + \lambda}{1 + e^{j\phi}\lambda}\right)^{-k} e^{j\phi n} d\phi$$

$$= \sum_{k=\infty}^{\infty} x(k) q(\lambda; n, k) \qquad (10)$$

with the definition of the interpolation function q

$$q(\lambda; n, k) = \frac{1}{2\pi} \int_{<2\pi>} \left(\frac{e^{j\theta} + \lambda}{1 + e^{j\theta}\lambda}\right)^{-k} e^{j\theta n} d\theta = F_n^{-1}\left\{\left(\frac{e^{j\theta} + \lambda}{1 + e^{j\theta}\lambda}\right)^{-k}\right\} \qquad (11)$$

and F_n^{-1} denoting the inverse Fourier transformation to the n-domain. More specifically,

$q(\lambda; n, 0) = \delta(n);$

$q(\lambda; -, k) =$ impulse response of an kth order all-pass, k>0,

$q(\lambda; n, k) = q(\lambda; -n, -k)$

$q(\lambda; n, k) = 0$, if $n \cdot k < 0$ or ($k = 0$ and $n \neq 0$).

In matrix notation (omitting λ from the notation for this specific case) equation (7) can be written as:

$$\begin{pmatrix} \vdots \\ x_m(-n) \\ \vdots \\ x_m(-1) \\ x(0) \\ x(1) \\ \vdots \\ x_m(n) \\ \vdots \end{pmatrix} = \mathrm{B} \cdot \begin{pmatrix} x_m^0(-N_1) \\ \vdots \\ x_m^0(-1) \\ x_m^0(0) \\ x_m^0(1) \\ \vdots \\ x_m^0(N_2) \end{pmatrix}$$

$$\begin{pmatrix} \vdots \\ x_m(-n) \\ \vdots \\ x_m(-1) \\ x(0) \\ x(1) \\ \vdots \\ x_m(n) \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots & & \vdots & & & & \\ q(n,N_1) & \cdots & q(n,1) & & & & \\ \vdots & & \vdots & & & & \\ q(1,N_1) & \cdots & q(1,1) & & & & \\ q(0,N_1) & \cdots & q(0,1) & 1 & q(0,1) & \cdots & q(0,N_2) \\ & & & & q(1,1) & \cdots & q(1,N_2) \\ & & & & \vdots & & \vdots \\ & & & & q(n,1) & \cdots & q(n,N_2) \\ & & & & \vdots & & \vdots \end{pmatrix} \begin{pmatrix} x_m^0(-N_1) \\ \vdots \\ x_m^0(-1) \\ x_m^0(0) \\ x_m^0(1) \\ \vdots \\ x_m^0(N_2) \end{pmatrix} \quad (12)$$

i.e. column-wise the impulse responses of the cascaded all-pass filters appear. In practice, a truncated (windowed) warped signal $\tilde{x}$ will be used for further processing. Assuming that the part of $\tilde{x}$ shall consider ranges from $-M_1$ to M_2 and that $M_1 \approx M_2 > 0$ and $N_1 \approx N_2$. Then, approximately half of the matrix equals zero. For positive λ, the support of the truncated $\tilde{x}$ will effectively be shorter than that of x.

The rows of the matrix correspond to the (truncated) impulse response of the filters described in embodiment 2.

It should be noted that the above-mentioned embodiments illustrate rather than limit the invention, and that those skilled in the art will be able to design many alternative embodiments without departing from the scope of the appended claims. In the claims, any reference signs placed between parentheses shall not be construed as limiting the claim. The word 'comprising' does not exclude the presence of other elements or steps than those listed in a claim. The invention can be implemented by means of hardware comprising several distinct elements, and by means of a suitably programmed computer. In a device claim

enumerating several means, several of these means can be embodied by one and the same item of hardware. The mere fact that certain measures are recited in mutually different dependent claims does not indicate that a combination of these measures cannot be used to advantage.